ABSTRACT

Data with good quality is a valuable asset for a company. Data can be processed into information to help companies improve decision making. Over time, the data owned by the company will increase more and more. However, more data can increase the tendency to arise problems about data quality. Thus, good data management is important to maintain data quality in meeting company standards. One of the efforts that can be done to clean up is clean the data from errors, inaccuracies, duplication, format differences, or other anomalies. This research will discuss the application of data cleansing using the Analytics Canvas method to customer account data at telecommunication companies. Data cleansing will be applied to customer account datasets and payment bill datasets totaling tens of millions of lines using the Apache Spark SparkSQL module to obtain good query processing performance. There are three stages in this study, namely preprocessing stage, processing stage, and validation stage. In addition, this study also reviewed the performance of Apache Spark in processing queries. In this study, Spark and Oracle's performance will be compared based on query processing time. Both will be tested on data cleansing queries and validation stage queries. After the application of data cleansing is complete, results are obtained in the form of quality datasets that contain customer accounts with the amount of 30% of the total initial dataset. Another study result was that there was a difference in query processing time on both tools. Apache Spark is rated better because it has a relatively faster query processing time than Oracle Database. It can be concluded that Oracle is more reliable when it comes to storing complex data models than in conducting data analysis. For future research, this research can be used as a basis for query optimization needs so that the most effective queries can be obtained with the fastest processing time.

Keywords—customer account, data cleansing, Apache Spark, query, caomparison